NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

DETEÇÃO DE FRAUDE EM ACIDENTES DE TRABALHO NO MUNICÍPIO DE OEIRAS

Maria Miguel Matos Menezes de Sequeira

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DETEÇÃO DE FRAUDE EM ACIDENTES DE TRABALHO NO MUNICÍPIO DE OEIRAS

por

Maria Miguel Matos Menezes de Sequeira

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação, especialização em Análise e Gestão de Risco

Orientador: Rui Alexandre Henriques Gonçalves

fevereiro 2020

AGRADECIMENTOS

Ao meu orientador, professor Rui Gonçalves por todo o apoio, tempo dispensado e conhecimento transmitido, durante todo o trabalho.

À Unidade de Saúde e Segurança no Trabalho pela disponibilização dos dados referentes aos acidentes de trabalho na CMO.

Aos meus pais por todos os ensinamento e oportunidades que me foram dando ao longo dos anos e que contribuíram decisivamente no desenvolvimento da pessoa que sou hoje.

À minha família e em especial aos meus avós que foram um complemento essencial no meu crescimento como pessoa e estudante.

Ao Zé pelo apoio, paciência e companhia dada nesta fase tão importante da minha vida.

Às minhas amigas de infância e de percurso que ao longo de 20 anos de existência consciente me tornaram no que sou hoje e me apoiaram sem perguntas aderentes. Em especial à Jessica pela companhia no desenvolvimento da dissertação.

RESUMO

Os acidentes de trabalho em Portugal têm vindo a aumentar de forma gradual. Paralelo a este fenómeno, o conceito de fraude é cada vez mais abordado por ser um tema que preocupa as organizações públicas e privadas, devido à sua difícil deteção e prevenção. A utilização de técnicas para detetá-las significa identificar tendências gerais de comportamentos suspeitos ou possíveis de fraude. É neste contexto que se insere esta tese, que apresenta um modelo preditivo capaz de prever a ocorrência de fraude em acidentes de trabalho no município de Oeiras. De forma a cumprir o objetivo foi recolhido o histórico dos acidentes ocorridos nos últimos cinco anos na organização e aplicado os algoritmos estudados na revisão de literatura. Através da análise e da comparação dos modelos construídos, é possível concluir que a sua eficácia ficou aquém do esperado. No entanto, reproduzindo a mesma análise para uma base de dados segregada por apenas uma categoria de lesão, foram obtidos melhores resultados.

PALAVRAS CHAVE

Acidente de Trabalho; Fraude; Data Mining; Município de Oeiras

ABSTRACT

Occupational accidents in Portugal have been gradually increasing. Analogous to this increase, the concept of fraud is increasingly addressed because it is a topic that concerns public and private organizations, due to its difficult detection and prevention. Using techniques to detect them requires identification of general trends in suspicious or possible fraud behaviour. It is in this context that this thesis is inserted, presenting a predictive model capable of predicting the occurrence of fraud in work accidents in the municipality of Oeiras. In order to fulfil the objective, the accidents that occurred in the last five years in the organization were collected and the algorithms studied in the literature review were applied. Through the analysis and comparison of the built models it is possible to conclude that its effectiveness was below the expected. However, reproducing the same analysis for a database segregated by only one category of injury, better results were obtained.

KEYWORDS

Occupational accident; Fraud; Data Mining; Municipality of Oeiras

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E ACRÓNIMOS

CMO - Câmara Municipal de Oeiras

ROC – Relative Operating Characteristic

RN – Rede(s) Neuronal(nais)

1. INTRODUÇÃO

1.1. ANTECEDENTES E QUADRO TEÓRICO

Como base deste estudo apresentaremos o conceito de risco operacional e a sua gestão nas organizações e uma noção inicial de fraude.

Risco operacional é inerente a todos os produtos, serviços e atividades empresariais e a gestão eficaz do risco operacional sempre foi um elemento fundamental de um programa de gestão de uma empresa. Como resultado, uma boa gestão dos riscos operacionais é um reflexo da eficácia da administração do Conselho e da empresa na administração dos seus portfólios de produtos, serviços e atividades.

A gestão de risco abrange o processo de identificação de riscos e sua medição (sempre que possível), garantindo um capital efetivo, programa de gestão e monitorização. Monitorando as exposições a riscos e necessidades de capital correspondentes, as empresas poderão continuamente tomar medidas para controlar ou mitigar os riscos. (ABI, 2011)

As fraudes internas e externas são incluídas na categoria de "risco operacional" pelo Consórcio de Seguro de Risco Operacional e definidas como má conduta intencional e atividades não autorizadas por partes internas ou externas, respetivamente. (Patel, 2010). Neste estudo, iremos ter como referência apenas a fraude interna.

Fraude interna é o risco de perda financeira, material ou de reputação inesperada como resultado de ações fraudulentas de pessoas internas à empresa. É uma categoria de risco reconhecida nas estruturas regulatórias em todo o mundo (padrões de Basileia II). A definição de Basileia II é mais especificamente 1) perdas por atos de um tipo destinado a fraudar, 2) apropriar-se de propriedade inadequada e/ou 3) burlar regulamentos, a lei ou a política da empresa, excluindo eventos de diversidade/discriminação, que envolvam pelo menos uma parte interna (Basel Committee on Banking Supervision, 2001).

Atualmente, os métodos de deteção de fraude podem ser divididos em quatro categorias: regras de negócios, auditoria, redes e modelos estatísticos e *data mining* (DM) (Shao & Pound, 1999). O uso da *data mining* deve-se à sua eficiência financeira que encontra evidências de fraude através da aplicação de algoritmos matemáticos nos dados disponíveis (Phua, Lee, Smith, & Gayler, 2010).

1.2. RELEVÂNCIA DO ESTUDO

Nesta secção, pretendemos esclarecer as razões pelas quais devemos preocupar-nos com fraudes e seus métodos de deteção, considerando a perspetiva das instituições. Será usada como caso de estudo a Câmara Municipal de Oeiras (CMO), particularmente a área dos acidentes de trabalho reportados pelos colaboradores.

De acordo com a Base de Dados de Portugal Contemporâneo a população na maioria dos concelhos tem vindo aumentar desde 2001 até ao presente. Particularmente no concelho de Oeiras, a população residente em 2017 teve um aumento de 7,63% em relação ao registado em 2001 (PorData, 2017).

O aumento populacional e necessidade constante de melhoria nos serviços municipais levou a CMO a aumentar em 663,9 mil euros a sua despesa com o pessoal (Comunicação do Sr. Presidente à Assembleia Municipal, 2019)

De acordo com o Gabinete de Estratégia e Planeamento (GEP), em 2016 registaram-se 207.567 acidentes de trabalho dos quais 138 foram mortais em Portugal. Dos acidentes ocorridos no ano de 2016 houve um período médio de 37,4 dias de afastamento (GEP, 2015). No gráfico abaixo (Figura 1), é possível observar a evolução do número de acidentes de trabalho em Portugal de 2007 a 2017. Apesar da descida acentuada de 2008 a 2012, estes têm vindo a aumentar de forma gradual.

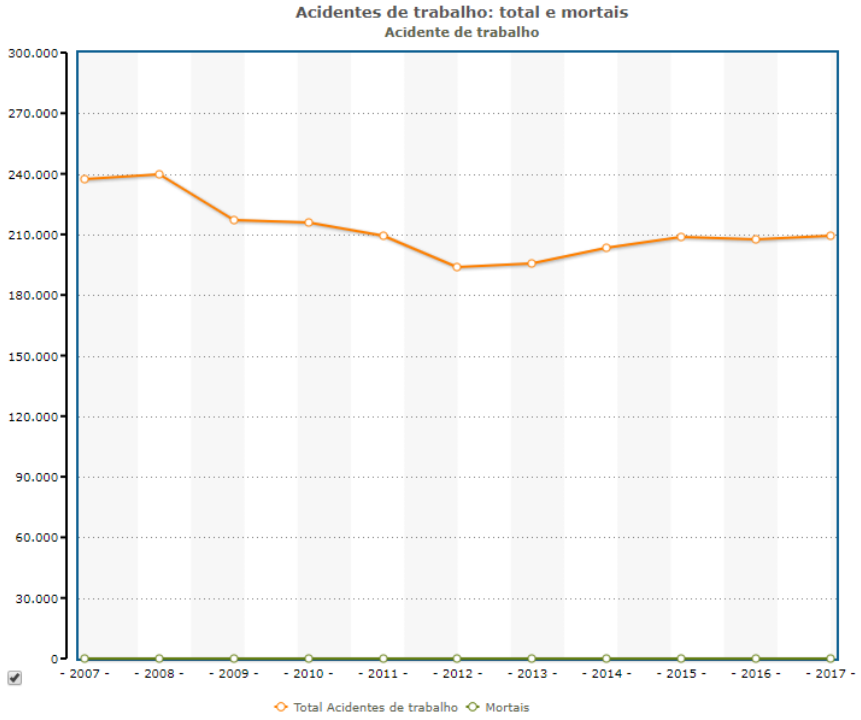

Figura 1 – Acidentes de trabalho em Portugal
Fonte: Acidentes de Trabalho (PorData, 2017)

Em suma, é possível, tendo em conta as informações providenciadas acima, afirmar que tem havido um aumento de populacional no concelho, do número de trabalhadores na CMO e do número de acidentes de trabalho. Podemos concluir que a CMO tem, ao longo dos anos, vindo a estar mais exposta à ocorrência de fraude nos acidentes de trabalho. A sua análise e prévia deteção tornam-se cada vez mais emergentes.

1.3. IDENTIFICAÇÃO DO PROBLEMA E OBJETIVOS

A vulnerabilidade à fraude nos acidentes de trabalho é agora óbvia. Uma estratégia agressiva deve ser posta em prática, sem interromper o relacionamento com os trabalhadores, investigando-os com muita frequência ou profundidade. As empresas precisam de adotar novas metodologias para detetar fraudes e reduzir o tempo necessário para processos manuais devido a reclamações fraudulentas (Phua *et al*., 2010). As perdas indubitavelmente severas devidas a esquemas de fraude devem ser motivo suficiente para estabelecer métodos mais eficientes de deteção e a sua prevenção, com o objetivo de evitar futuras alegações suspeitas antes mesmo do registo.

Constitui objetivo principal deste estudo a realização de um modelo de fraude para os acidentes de trabalho ocorridos no Município de Oeiras. A análise da sinistralidade laboral será executada com a parceria da Divisão de Promoção Socioprofissional e a Unidade de Segurança e Saúde no Trabalho.

Recolhendo os dados reportados nos últimos 5 anos (2013-2018) à unidade e todos os parâmetros que poderão estar relacionados com o acidente é objetivo:

1. Realizar uma análise estatística através de métodos não supervisionados com base nas variáveis;

2. Criação da variável *target* do estudo, possível existência de fraude, com base na análise realizada no ponto anterior;

3. Construir um modelo preditivo capaz de prever a ocorrência de fraude.

Na próxima secção, explicaremos como pretendemos atingir as metas definidas, identificando processos e procedimentos destinados a obter as informações necessárias para criar o modelo.

2. REVISÃO DE LITERATURA

2.1. RISCO OPERACIONAL

Risco é um evento incerto ou uma condição que, ocorrendo, tem um efeito positivo ou negativo sobre um projeto (PMI, 2008).

A gestão de risco é o processo de gestão, planeamento, avaliação e controlo dos processos de uma empresa, com o objetivo de eliminar o possível risco da empresa e melhorar seu desenvolvimento, lucro e resultados financeiros. As instituições devem ter modelos de gestão de risco bem definidos e que devem incluir avaliação, análise e eliminação de risco. Como visto na Figura 2, a gestão de riscos deve abranger todos os processos, relatórios, estratégias e procedimentos para identificar, monitorizar, medir, gerir e relatar os riscos de forma contínua. A cultura de risco pode ser entendida como as normas e tradições de comportamento de indivíduos e grupos dentro de uma organização que determinam a maneira através da qual eles identificam, entendem, discutem e agem sobre os riscos que a organização confronta e assume. De facto, a apetência ao risco define o volume de risco total que a organização aceita reter (Kaļiņina & Voronova, 2014).

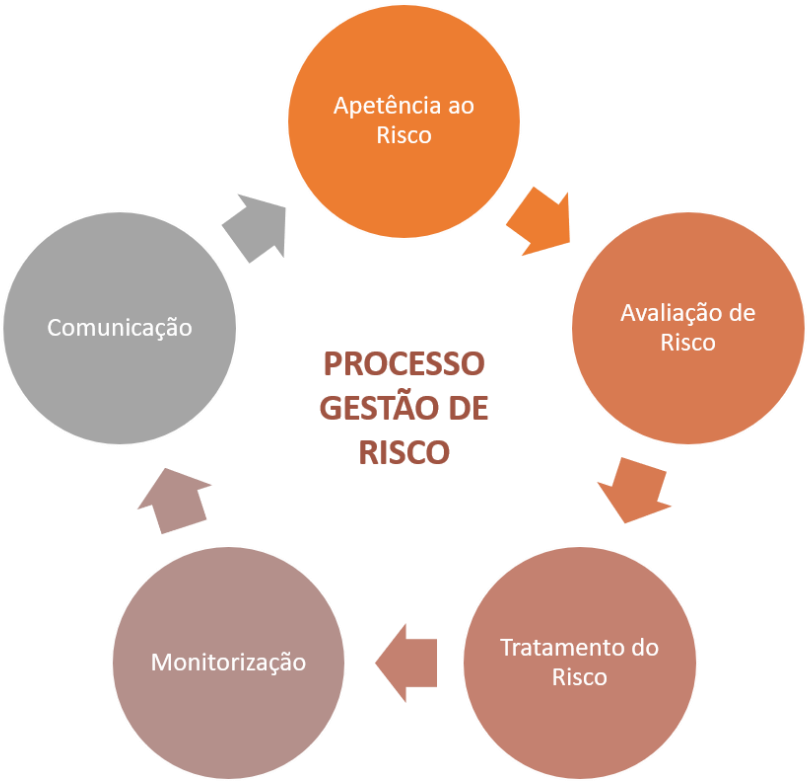

Figura 2 - Processo de Gestão de Risco
Fonte: Risk Management Process (Marel, 2017)

O risco pode ser categorizado pela tipologia da exposição em várias categorias, como risco de mercado e risco de crédito (ambos dentro do risco financeiro), risco de liquidez, risco legal e regulatório, risco estratégico e comercial, risco operacional e risco de reputação (European Parliament and European Council, 2009). O risco operacional é definido como o risco de perdas decorrentes de processos inadequados ou falhas internas, pessoas e sistemas ou de eventos externos. O risco operacional inclui riscos legais, mas exclui o risco de reputação (European Banking Authority, 2016).

Os riscos operacionais são difíceis de definir devido ao amplo espectro de potenciais perdas que cobrem. De acordo com o segmento em que a empresa atua, isso pode estar sujeito a vários riscos operacionais inerentes ao negócio. A Tabela 1 - Risco Operacional, apresentada abaixo, lista, na visão da *British Banking Association* (BBA), os principais aspetos do risco operacional a serem considerados e controlados. Os aspetos de risco operacional interno incluem pessoas, processos e sistemas, enquanto o risco operacional externo apresenta categorias e subcategorias aceites pelos investigadores (Mackenzie, Kimura, Kerr, Mackenzie, & Lima, 2010).

Risco Interno			**Risco Externo**	
Pessoas	Processos	Sistemas		
Faude	Erros de contabilidade	Qualidade dos dados	Externo	Legal
Erro do trabalhador	Rico de capacidade	Erros de programação		Lavagem de dinheiro
Perda de trabalhadores	Risco de contrato	Problemas de seguro		Problemas políticos
Empréstimo de trabalhadores	Erros de pagamento	Capacidade do sistema		Regulação
Saúde e segurança	Erros de validação	Compatibilidade do sistema		Taxas
Falta de conhecimento e técnica	Erros de trasação	Falhas no sistema	Físico	Fogo, Roubo, terrorismo
				Desastres naturais
				Segurança física

Tabela 1 - Risco Operacional
Fonte: Operational Risk (Mackenzie et al., 2010)

No próximo capítulo, iremos explorar com maior detalhe um dos aspetos listados pela BBA como risco operacional: a fraude.

2.2. ACIDENTES DE TRABALHO

Este capítulo tem como objetivo aprofundar alguns conceitos que serão abordados ao longo do trabalho desenvolvido e realizar um enquadramento em alguns desses temas, como os acidentes e a sua ocorrência no exercício profissional. Deste modo, começa-se por fazer um breve enquadramento histórico para o entendimento sobre a ocorrência dos acidentes, assim como a evolução destes.

Acidente é uma palavra com uma rica linhagem filosófica e uma interessante história etimológica. Aristóteles usou esta palavra para definir características não essenciais ou extrínsecas (Denominations & Church, n.d.). Assim, pessoas e coisas tinham qualidades substanciais e acidentais. Duas pernas não eram uma qualidade substancial de humanos, eram

uma qualidade acidental porque alguns animais também andam em duas pernas (Hermann, Dr, & Guamieri, 1996).

Cerca do século XIV, no auge da influência normanda, os ingleses começaram a usar outro significado da palavra (que perdura até aos dias de hoje): acontecer por acaso; um infortúnio; um evento que acontece sem previsão ou expectativa. Este significado está associado à palavra francesa *accidence*, que pode ser uma corrupção do verbo latino *accidere*, que significa cair ou cair para (Hermann *et al*., 1996).

O significado moderno da palavra "acidente" apareceu na primeira década do século XV. O seu sucesso foi compreensível, pois a nova palavra transmitia uma mistura de ideias: lesão, perda de propriedade, eventos inesperados e resultados não intencionais. Como o mundo medieval tinha poucas ferramentas científicas para investigar a causalidade, a palavra associou-se a um ato divino, à frase ato de Deus (Hermann *et al.*, 1996).

Apenas em 1931 foi desenvolvido o primeiro modelo de causa, bem como estimativa de custo de acidentes, estendendo essas análises nas próprias empresas participantes através de entrevistas com membros do *staff* dos serviços de administração e produção. No livro *Accident Prevention*, Heinrich, em 1959, refere que os acidentes de trabalho com ou sem lesões são devidos a vários fatores como personalidade do empregado, prática de atos inseguros, existência de condições inseguras nos locais de trabalho, entre outros (Navarro, 2017).

Se for considerado que os acidentes de trabalho resultam do avanço tecnológico e dos processos de industrialização, então os principais responsáveis são a ciência, a técnica e o progresso, em vez dos empregadores (*in* Pinto, 1996: 105). Ulrich Beck (1992) e Charles Perrow (1999) corroboram a ideia de que a ciência e a tecnologia acarretam novas formas de risco para a modernidade, passíveis de originar acidentes ou efeitos devastadores (Mendes, 2015).

Na República Portuguesa, a Lei 98/2009, através do art.º 8.º, define os seguintes conceitos:

- **Acidente de trabalho** - aquele que se verifique no local e no tempo de trabalho e produza direta ou indiretamente lesão corporal, perturbação funcional ou doença de que resulte redução na capacidade de trabalho ou de ganho ou a morte

- **Local de trabalho** - todo o lugar em que o trabalhador se encontra ou deva dirigir-se em virtude do seu trabalho e em que esteja, direta ou indiretamente, sujeito ao controlo do empregador

- **Tempo de trabalho** - além do período normal de trabalho, o que precede o seu início, em atos de preparação ou com ele relacionados, e o que se lhe segue, em atos também com ele relacionados, e ainda as interrupções normais ou forçosas de trabalho.

(Estado Português, 2009)

O conceito de acidente de trabalho pressupõe a verificação cumulativa dos seguintes elementos:

- **Acidente**;

- **de trabalho**: ocorrido – no local de trabalho OU no tempo de trabalho;
- verificando-se: um nexo de causalidade entre:
    - o trabalho e o evento;
    - o evento e a lesão;
    - a lesão e a incapacidade

(Beatriz, 2019)

2.3. FRAUDE

Quando falamos em fraude, não falamos apenas de fraude fiscal, apesar de esta ser a única comtemplada na lei. A fraude engloba várias situações, tendencialmente intencionais, em que os cidadãos ou instituições causam direta ou indiretamente danos económicos e/ou sociais. Para que seja considerado legalmente fraude é necessário que existam danos, geralmente económicos, para a vítima. Existem quatro elementos gerais que devem ocorrer para que esta seja considerada:

1. Declaração material falsa;
2. Conhecimento de que a declaração era falsa quando foi comunicada;
3. Confiança na declaração falsa por parte da vítima;
4. Danos consequentes.

(Pimenta, 2009)

Figura 3 - Ciclo de Fraude
Fonte: Fraud Cycle (Baesens, Vlasselaer, & Verbeke, 2015)

A Figura 3 representa o ciclo de fraude com os seus quatro elementos base:

- **Deteção:** Aplicar modelos de deteção em novas observações e atribuir um risco de fraude a cada uma delas;
- **Investigação:** Uma pessoa especialista é geralmente encarregue de investigar casos suspeitos e sinalizados, dada a subtileza e complexidade envolvidas neste processo;
- **Confirmação:** Determinação dos verdadeiros casos de fraude, com possibilidade de investigação no campo;
- **Prevenção:** Impedir que uma fraude seja cometida no futuro. Isto pode resultar em deteção de fraude antes mesmo de uma pessoa saber que irá cometer a mesma.

Os casos recentemente detetados devem ser adicionados à base de dados de que contém os históricos de fraude, este é utilizado para aprender ou induzir um modelo de deteção. O modelo de deteção de fraude não deve ser refeito sempre que existe um novo caso confirmado. No entanto, é recomendável uma atualização regular do modelo, dada a natureza dinâmica da fraude. A frequência necessária para refazer ou atualizar o modelo de deteção depende de vários fatores:

- Volatilidade do comportamento fraudulento;
- Poder de deteção do modelo atual, relacionado com a volatilidade do comportamento de fraude;

- Quantidade de casos confirmados (semelhantes) já disponíveis na base de dados;
- Percentagem de casos que estão a ser confirmados como verdadeiros casos de fraude;
- Esforço necessário para refazer o modelo.

(Baesens *et al.*, 2015)

Geralmente, a fraude ocorre devido a uma combinação de oportunidade, pressão e racionalização. Uma oportunidade surge, o indivíduo sente que o ato não está totalmente errado e isso pressiona-o a cometer a fraude (Glover & Flagg, 1999).

Curiosamente, estudos mostraram que a remoção da pressão não é suficiente para eliminar a fraude. Além disso, o primeiro ato de fraude exige mais racionalização do que o segundo, e assim por diante. Contudo, à medida que se torna mais fácil justificar, os atos ocorrem com mais frequência e os montantes envolvidos aumentam de valor. Isso significa que, deixada ao abandono, a fraude continuará e as perdas só irão aumentar. A fraude, em última análise, é alimentada pela ganância e a ganância nunca acaba. (Glover & Flagg, 1999)

As responsabilidades relacionadas com a prevenção de fraude dentro de uma organização estão divididas entre o conselho executivo, o comité de auditoria e a auditoria interna. Em primeiro lugar, o executivo tem a responsabilidade final de implementar os mecanismos de deteção e prevenção de uma fraude desde o início. Em segundo lugar, o comité de auditoria tem o papel de supervisionar a gestão de riscos de fraude, monitorizando ativamente os esforços do conselho executivo contra a prática de fraude. Em terceiro lugar, a auditoria representa uma linha eficiente de defesa contra a fraude, desempenhando um papel tanto na monitorização de riscos quanto na fraude, prevenção e deteção. A auditoria interna constitui uma ferramenta à disposição do comité de auditoria, o único capaz de avaliar independentemente os riscos de fraude e as medidas antifraude implementadas pelo conselho executivo (Petraşcu & Tieanu, 2014)

2.4. Fraude em Acidentes de Trabalho

A fraude na área dos acidentes de trabalho e compensação dos trabalhadores pode dividir-se em vários tipos:

- Fraude do empregador;
- Fraude do trabalhador;
- Fraude do profissional de saúde;
- Outra fraude - a fraude nem sempre é restrita às pessoas diretamente envolvidas, poderá ser cometida contra o sistema de compensação dos trabalhadores por alguém indiretamente envolvido na reivindicação.

O tipo de frade abordado neste trabalho é a fraude cometida pelo trabalhador. Existem diferentes tipos de fraude do trabalhador. Os exemplos podem incluir:

- Reivindicação por uma lesão que não ocorreu no trabalho;
- Ausência de notificação de retorno ao trabalho;
- Não divulgação de múltiplas reivindicações relacionadas ao mesmo prejuízo;
- Reivindicações falsas ou exageradas de viagens e/ou despesas;
- Falsificação de atestados médicos;
- Fornecer informações falsas ou enganosas em relação a uma reivindicação.

Os indicadores de fraude podem não estabelecer conclusivamente a existência de fraude. Pode ser necessária uma investigação adicional para determinar se ocorreu uma fraude.

(SIRA, 2018)

2.5. TÉCNICAS DE DETEÇÃO DA FRAUDE

A partilha de conhecimento no tema "deteção de fraude" é limitada, logo o desenvolvimento de novos métodos é mais lento e difícil. A partilha de técnicas de deteção de fraude em grande detalhe ao domínio público é feita deliberadamente, pois esse conhecimento irá fornecer aos potenciais criminosos as informações necessárias para evitar a deteção (Bolton & Hand, 2002).

O objetivo do uso de técnicas para detetar fraudes é identificar tendências gerais de comportamentos suspeitos de fraude (Wang, 2010). Considerando o estudo da fraude, as técnicas que utilizam estatística e *data mining* são o futuro dessa área, pois fornecem tecnologias efetivas para a deteção de fraudes (Bolton & Hand, 2002).

As ferramentas estatísticas para deteção de fraude são muitas e variadas, pois os dados podem ser diferentes em tamanho e tipo, mas existem conceitos comuns. Tais ferramentas são essencialmente baseadas na comparação dos dados observados com os valores esperados. Os métodos estatísticos de deteção de fraude podem ser supervisionados ou não supervisionados (Bolton & Hand, 2002).

Os métodos não supervisionados baseiam-se na pesquisa de *outliers* no conjunto de dados, ou seja, nas observações que diferem da norma. De facto, não é certo que um caso de fraude seja exposto, mas a análise deste tipo de método visa alertar para o facto de uma observação ser anormal e necessitar de ser investigada com mais detalhe (Bolton & Hand, 2002). Este tipo de método é utilizado quando não há conjuntos anteriores de observações legítimas e fraudulentas. É traçada uma linha de distribuição que representa o comportamento normal e, de seguida, o objetivo é detetar observações que apresentem um grande afastamento dessa

norma (Kou, Lu, Sirwongwattana & Huang, 2004). Exemplos: análise de dígitos utilizando a lei de Benford e criação de pequenos grupos.

A lei de Benford (1938) é baseada em estudos de tabelas de logaritmos, que aponta diferenças entre as frequências observadas entre os primeiros algarismos significativos. A lei de Benford afirma que a distribuição dos primeiros dígitos significativos de números extraídos de uma ampla variedade de distribuições aleatórias terá assintoticamente uma certa forma (Silva, Korzenowski & Vaccaro, 2014). Em relação aos *clusters*, o método consiste no agrupamento de dados com as mesmas características em pequenos grupos (“*clusters”)*. Dentro desses grupos, são identificadas entidades com recursos que diferem mais dos outros em diferentes *clusters*. As características nesse caso significam a quantidade de valores estatísticos que são comparados juntos (Hawlova, 2013).

Nos métodos supervisionados, amostras de registos fraudulentos e não fraudulentos são usadas para construir o modelo. Nesse sentido, é necessário ter exemplos de ambas as classes e só pode ser usado para detetar fraudes que ocorreram anteriormente (Bolton & Hand, 2002). O método é baseado em exemplos do passado que são usados para preparar e “treinar” o modelo estatístico que calcula a probabilidade de ocorrência de fraude (Hawlova, 2013). A técnica de regressão, que consiste na construção de um modelo de regressão que tenta descobrir a relação entre variáveis independentes e uma variável alvo, é normalmente usada para a deteção de fraudes corporativas, como é o caso deste estudo em particular (Ngai, Hu, Wong, Chen & Sun, 2011).

Data mining é definida como o processo de descoberta de padrões em grande quantidade de dados (Witten, Frank & Hall, 2011). Pode ser introduzida como o processo de adoção de um conjunto de ferramentas e procedimentos de análise de dados para identificar padrões e relacionamentos nos dados para os resumir em informações úteis (Fahmi, Hamdy & Nagati, 2016). *Data mining* é mais útil num cenário de análise exploratória em que não há noções predeterminadas sobre o que constituirá um resultado ''interessante'' (Kirkos, Spathis & Manolopoulos, 2007).

Os hábitos de um indivíduo tendem a exibir uma natureza específica e esse perfil comportamental pode ser usado para interpretar um conjunto de padrões. A variação de certos padrões implica uma possível ameaça ao sistema (Fahmi *et al*., 2016). A capacidade da *data mining*, como reconhecer padrões, classificar dados e identificar modelos de dados poderá ser utilizada para identificar fraude nos acidentes de trabalho.

Embora os modelos estatísticos para deteção de fraude possam ser categorizados como supervisionados ou não supervisionados, as áreas de aplicação da deteção de fraude são muito diversas e, de acordo com a variedade e quantidade de dados disponíveis, é necessário escolher a ferramenta de deteção de fraude mais adequada (Bolton & Hand, 2002). Algumas técnicas sobre DM serão abordadas e explicadas em mais detalhe nos seguintes subcapítulos.

2.5.1. Árvores de Decisão

Árvores de Decisão (AD) trata-se de um modelo de classificação baseado em dados que consiste numa estrutura em árvore, onde cada nó representa um teste num atributo e cada ramificação representa um resultado desse mesmo teste. Nesse sentido, as observações são divididas em subgrupos mutuamente exclusivos (Kirkos *et al.*, 2007). A amostra é sucessivamente dividida em subconjuntos até que nenhuma outra divisão possa produzir diferenças estatisticamente significativas.

Para este tipo de modelo de classificação, é necessário um conjunto de exemplos estruturados com variáveis não categóricas - as entradas - e uma variável categórica - a saída. Em seguida, o objetivo é encontrar um modelo, denominada árvore de decisão, que possa classificar corretamente a que categoria pertencem os dados não categóricos apresentados com novos valores. Relacionada com o tipo de variáveis de entrada e saída, a primeira pode ser contínua ou discreta enquanto o tipo de saída é discreto e, em geral, do tipo binário. Isso significa que a variável assume os valores 1 ou 0, o que representa se pertence ou não a uma categoria, respetivamente (Song & Lu, 2015).

Quando o tamanho da amostra é suficientemente grande, os dados do estudo podem ser divididos em conjuntos de dados de treino e validação. Utilizando o conjunto de dados de treino para criar um modelo de árvore de decisão e um conjunto de dados de validação para decidir sobre o tamanho de árvore apropriado necessário para alcançar o modelo final ideal (Song & Lu, 2015). O uso de árvores de decisão fornece uma maneira significativa de representar o conhecimento adquirido e facilita a extração de regras de classificação *if-then* (Kirkos *et al.*, 2007).

As árvores de decisão podem ser utilizadas para vários propósitos na análise de fraude:

- Seleção de variáveis, pois as variáveis que ocorrem no topo da árvore são mais preditivas;
- Calcular o poder preditivo de uma certa variável/característica;
- Segmentação para melhoria de outro modelo, isto é, construir uma árvore com dois ou três níveis de profundidade, à medida que a química das segmentações dispensa os modelos de regressão logística do segundo estágio para refinamento adicional;
- Modelo final de fraude analítica a ser usado diretamente no ambiente de negócios.

(Baesens *et al.*, 2015)

A desvantagem deste modelo é a sua dependência na amostra utilizada para a construção das árvores. Uma pequena variação na amostra subjacente pode produzir uma árvore totalmente diferente (Baesens *et al*., 2015).

2.5.2. Regressão Logística

A regressão é uma técnica de *data mining* utilizada para ajustar uma equação a um conjunto de dados. A forma mais simples de regressão, regressão linear, utiliza a fórmula de uma linha reta (y = mx + b) e determina os valores apropriados para m e b para prever o valor de y com base num determinado valor de x (Gupta, 2015). Este modelo define uma relação linear entre a variável dependente e uma variável independente onde o fator que é previsto (o fator para o qual a equação resolve) é a variável dependente e o fator que é utilizado para prever o valor da variável dependente é a variável independente. Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se modelo de regressão linear múltipla (Figura 4) (Henriques, 2011).

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_1 X_{2_i} + ... + \beta_k X_{k_i} + e_i$$

Figura 4 - Equação de Regressão Linear Múltipla
Fonte: Análise de Regressão Linear Simples e Múltipla (Henriques, 2011)

Os X são as variáveis independentes e Y é a variável dependente. O índice i representa o número da observação (linha). Os β's são os coeficientes de regressão desconhecidos. As suas estimativas são representadas por b's, isto é, cada β representa o parâmetro desconhecido original (população), enquanto b é uma estimativa desse β. O εj é o erro (residual) da observação i (Boniface & Boniface, 2019).

A regressão logística é uma técnica de deteção de fraude supervisionada muito popular devido à sua simplicidade e bom desempenho. Assim como na regressão linear, uma vez que os parâmetros tenham sido estimados, eles podem ser avaliados de maneira direta, contribuindo assim para sua eficiência operacional (Baesens *et al.*, 2015).

2.5.3. Redes Neuronais

Uma rede neuronal (RN) é um conjunto de elementos interligados, unidades ou nós de processamento simples, cuja funcionalidade é baseada num neurónio animal. A capacidade de processamento da rede é armazenada nas forças ou pesos da conexão entre os nós obtidos por um processo de adaptação ou aprendizagem de um conjunto de padrões de treino (Cross, Harrison & Kennedy, 1995).

As RN são uma importante ferramenta de *data mining* utilizada para classificação e construção de *clusters*. É uma tentativa de construir uma máquina que imite as atividades cerebrais e que seja capaz de aprender. Se forem fornecidos às RN exemplos suficientes, ela poderá executar a classificação e até descobrir novas tendências ou padrões nos dados. As RN básicas são compostas por três camadas, entrada, saída e camada oculta. Cada camada tem um determinado número de nós. Os nós da camada de entrada estão conectados aos nós da camada oculta, os nós da camada oculta estão conectados aos nós da camada de saída como demonstrado na Figura 5 (Cilimkovic, 2010).

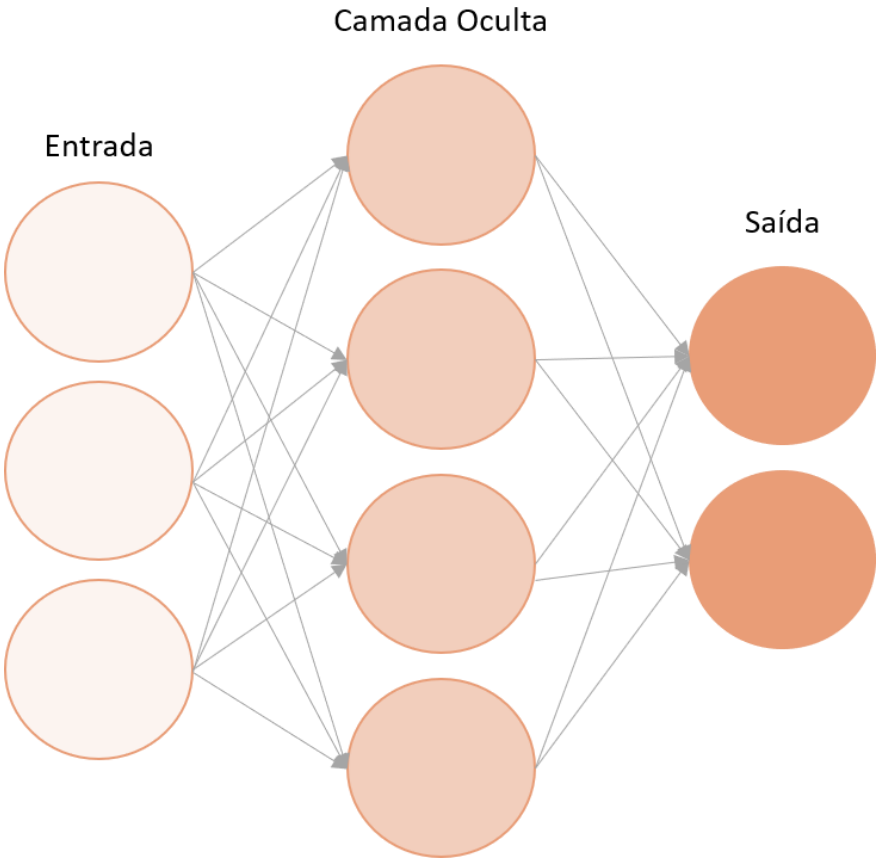

Figura 5 - Rede Neuronal
Fonte: Simple Neural Network (Cilimkovic, 2010)

- Entrada: representa as variáveis de entrada da RN;
- Camada Oculta: representa as influências de interação entre os neurónios. Uma RA pode ter mais de uma camada oculta ou não pode ter camadas ocultas;
- Saída: representa as variáveis de saída da RN.

As camadas podem ser divididas em quatro tipos, de acordo com o método de aprendizagem da RN: aprendizagem supervisionada, aprendizagem não supervisionado, aprendizagem associativa e aplicação de otimização.

1. Redes de aprendizagem supervisionadas: recebem exemplos de treino de problemas, incluindo valores de entrada e saída. Primeiro, a rede deve ser treinada e o peso da rede é ajustado com base na discrepância entre o valor de saída esperado e o valor calculado real;
2. Redes de aprendizagem não supervisionadas: contêm dados de entrada e não incluem os dados de saída esperados. Não há requisito de erro mínimo. O método de treino é baseado nos dados de entrada;

3. Redes de aprendizagem associada: aprende regras internas de memória associativa a partir de vários exemplos e as aplica a novos casos. Normalmente é utilizado para aquisição de dados e filtragem de ruído.

4. Redes de otimização: define os valores das variáveis de acordo com o problema. Sob a condição de satisfazer as restrições de *design*, tem como objetivo atingir o objetivo do problema de forma otimizada.

(Huang, 2013).

O tempo para treinar este método é identificado como a maior desvantagem. Este também exige conjuntos de amostras muito grandes para treinar o modelo com eficiência (Cilimkovic, 2010).

3. METODOLOGIA

Esta tese seguirá a metodologia de investigação científica para atingir o seu objetivo final. Esta metodologia de investigação requer que a motivação, o problema e os objetivos da investigação sejam claramente definidos.

Na primeira fase, foi importante apresentar o quadro teórico, relevância e objetivos deste estudo. Na introdução, é fornecida uma contextualização ampla do inerente problema da deteção de fraude partindo do conceito de risco operacional e finalizando no próprio conceito de fraude. Após o entendimento da importância deste estudo em acidentes de trabalho no concelho, foram definidos os objetivos. Para compreensão e análise dos métodos utilizados para deteção de fraude, é realizado um estudo dos modelos mais significativos.

3.1. RECOLHA E LIMPEZA DOS DADOS

No caso desta investigação, a informação foi recolhida na Unidade de Saúde e Segurança no Trabalho da CMO A informação dos acidentes de trabalho ocorridos nos últimos cinco anos no município foi fornecida em formato Excel. Na prática, esta informação é recolhida pela unidade através do preenchimento de uma ficha em formato de papel pelos sinistrados após o acidente. (Anexo I)

Consequente ao processo totalmente manual da passagem da informação física para um formato digital, existe um elevado risco de erro humano que não teremos em consideração neste estudo.

O ficheiro continha 1393 entradas com as seguintes colunas:

ID Acidente	N PROC	Categoria	Tipo	Causas	Vínculo
Serviço	Área Funcional	Data Nascimento	Sexo	Natureza Lesão	Data do AT
Hora do AT	Ficou internado?	Se Sim, onde?	Alta	Internamento	Local do AT
Assistência Médica	Circunstâncias da Ocorrência	Lesão	Tipo Lesão	Total Despesas	Grau Incapacidade Permanente
Percentagem incapacidade	Junta médica	Entidade	Data Alta	Observações	Dias de Ausência
Valor indemnização	Região	Zona atingida	Data pagamento	Atividade física	Agente Material
Desvio	Modalidade	*In itinere*	Ativo		

Foi necessário aplicar um processo de limpeza aos dados. Após uma análise inicial no próprio ficheiro Excel, foram tomadas as seguintes decisões:

1. As seguintes colunas não foram tidas em consideração pois estavam preenchidas com códigos internos dos quais não foi possível obter significado:
    - Atividade Física;
    - Agente Material;
    - Desvio;
    - Modalidade Lesão;
    - Tipo de Lesão.
2. Com uma percentagem superior a 80% de valores omissos, foram excluídas da base de dados as seguintes colunas:
    - Grau Incapacidade Permanente;
    - Percentagem incapacidade;
    - Data;
    - Junta Médica;
    - Entidade;
    - Percentagem Incapacidade;
    - Valor Indemnização.
3. Foram também excluídas as seguintes colunas e os motivos que levaram à sua exclusão são mencionados particularmente:
    - N PROC - consiste no número do processo, é utilizado como código interno pelo que não terá relevância para o nosso modelo de fraude;
    - Tipo (Acidente de Trabalho ou Reabertura) – visto 4% dos acidentes de trabalho serem reaberturas, esta variável foi removida pela sua homogeneidade;
    - Serviço – nome do serviço que o onde o trabalhador era empregado na altura do acidente. Foi excluída pois as nomenclaturas destes serviços são alteradas no início de cada mandato da CMO, deixando de ser possível relacionar os acidentes entre si;
    - Ficou internado? Se sim, onde? & Alta Internamento – foram excluídas pela sua homogeneidade, os casos de acidente com internamento foram abaixo dos 1% nestes últimos 5 anos;

- Observações – campo de texto aberto com possível inserção de múltiplas linhas de texto pelo que a sua heterogeneidade e conteúdo não se revelaram importantes para o estudo;

- Ativo – variável binária que indica se o processo do acidente ainda se encontra ativo ou não. Esta informação não é importante na deteção da fraude.

4. Foram eliminados da base de dados os acidentes que apresentavam a coluna Natureza em branco. 95,83% destes casos representavam reaberturas de processos ocorridos em anos anteriores. A coluna Natureza teve destaque neste ponto pois será necessária nas análises seguintes. Eliminando as 48 ocorrências com este critério, a base de dados ficou então com 1345 observações.

3.2. TRANSFORMAÇÃO DOS DADOS

Devido a alguma falta de critério no preenchimento das fichas de ocorrência, os dados continham muita informação não normalizada. A diversidade da informação inserida leva à necessidade de homogeneizar as colunas. Foram construídas as seguintes colunas com base nas já disponibilizadas:

A coluna Data de Nascimento e Data do AT continham mais de 800 valores diferentes, pelo que, para sua normalização, foram criadas duas colunas:

1. **Idade** – cálculo da idade do indivíduo na altura da ocorrência do acidente;

2. **Intervalo Idade** - tendo em conta a coluna anteriormente criada, foram agrupadas as ocorrências dos indivíduos com idades próximas, isto é, construção de uma variável qualitativa nominal com as seguintes categorias:

    - 20-30;
    - 31-40;
    - 41-50;
    - 51-60;
    - > 61.

A coluna Hora do AT continha 200 valores de hora diferentes pelo que a coluna Período foi criada para a sua homogeneização:

- **Período** - coluna qualitativa nominal com as seguintes categorias:

    - Manhã (8h01 às 12h00);
    - Tarde (12h01 às 18h00);
    - Noite (18h01 às 8h00).

A coluna Área Funcional continha 135 valores de áreas diferentes inseridas, pelo que a coluna Área foi criada para a sua homogeneização:

- **Área** – foram agrupadas as funções nas seguintes categorias:
    - Administrativo (Ex: Técnico Informático, Arquitetura e Serviços Sociais);
    - Limpeza/Higiene (Ex: Cantoneiro, Lavagem de Viaturas, Higiene Pública e Abastecimento);
    - Operacional (Ex: Jardinagem, Mecânico e Cozinheiro);
    - Segurança (Ex: Agente PM, Vigilante e Fiel de Armazém);
    - Transportes (Ex: Motorista, Condutor de Máquinas e Condutor de Transportes Públicos);

A coluna Local do AT continha 406 valores de locais diferentes inseridos pelo que a coluna Concelho foi criada para a sua homogeneização:

- **Concelho** - foram agrupados os locais nos seguintes concelhos do município:
    - Algés, Linda-a-Velha e Cruz Quebrada;
    - Barcarena;
    - Carnaxide e Queijas;
    - Oeiras e São Julião da Barra;
    - Porto Salvo;
    - Fora do concelho.

A coluna Região continha 48 valores de regiões do corpo afetadas diferentes e as circunstâncias, na coluna Circunstâncias descritas em modo texto livre pelo que a coluna Zona Atingida e No exercício laboral? Foram criadas para a sua homogeneização:

- **Zona Atingida** - foram agrupadas as partes do corpo atingidas no acidente nas seguintes categorias:
    - Cabeça;
    - Membro Inferiores;
    - Membros Superiores;
    - Tronco;
    - Múltiplas.
- **No exercício Laboral?** - foi classificado o acidente caso o mesmo tenha ocorrido aquando uma tarefa típica do profissional. Exemplo: caso um profissional de limpeza tenha tido um acidente a limpar as escadas é classificado como *Sim*, caso tenha ocorrido na deslocação para o trabalho é classificado como *Não*.

A coluna Natureza das Lesões continha 933 valores de naturezas de lesões diferentes inseridos pelo que a coluna Natureza foi criada para a sua homogeneização:

3. **Natureza** – foram classificadas as lesões nas seguintes categorias: Alergia, Algia, Amputação, Cervicalgia, Amputação, Conjuntivite, Contusão, Corpo Estranho, Corte, Danos Materiais, Dorsalgia, Edema, Entorse, Esmagamento, Ferida, Fratura, Hematoma, Hérnia, Inflamação, Irritação, Lesão, Lesão Muscular, Lombalgia, Luxação, Omalgia, Outro, Picada, Queimadura, Traumatismo.

Em consequência da construção das variáveis categóricas foram excluídas da análise as seguintes colunas por duplicação de informação:

- Data do AT;
- Hora do AT;
- Idade;
- Natureza das Lesões;
- Circunstâncias;
- Região;
- Local;
- Área Funcional;
- Data de Nacimento;
- Causas.

3.3. VARIÁVEIS QUALITATIVAS

Após a limpeza e transformação dos dados, foi realizada uma análise da distribuição dos acidentes pelas diferentes categorias. De forma a auxiliar a tarefa, foi utilizado um *software* de visualização de dados e construção de relatórios da Microsoft, o PowerBI. Esta ferramenta permite visualizar os dados e partilhar informações. É possível ligar-se a centenas de origens de dados, neste caso foi ligado ao ficheiro Excel.

Abaixo são mostrados alguns dos gráficos obtidos no relatório:

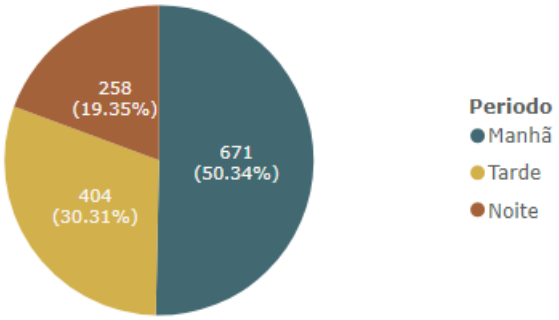

Figura 6 - Distribuição dos acidentes por Período do dia

Na figura 6, é possível verificar que o periodo onde ocorrem mais acidentes é o periodo da manhã, isto é, entre as 8h00 e as 12h00. Não existem factores relevantes para apontar uma

causa fiável mas, segundo Pablo Valdez, os componentes da atenção estão num nível baixo pela manhã (07h00 às 10h00), principalmente porque os ritmos circadianos atingem o seu ponto mais baixo a essa hora do dia e a inércia do sono também contribui para esse baixo nível de execução (Valdez, 2019).

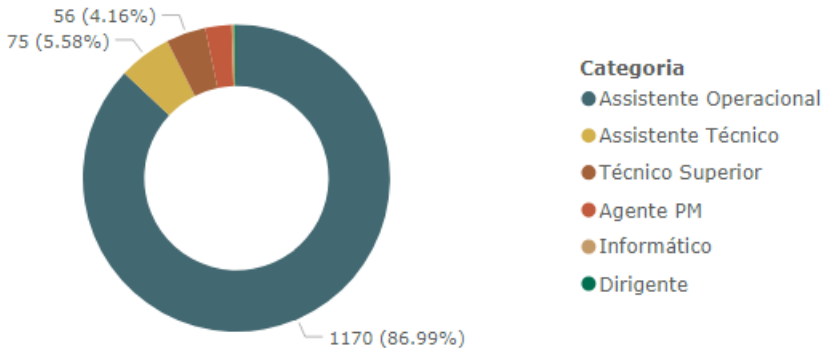

Figura 7 - Distribuição dos acidentes por Categoria

Através da figura 7, podemos verificar que 87% dos acidentes ocorrem em trabalhadores com a categoria Assistente Operacional. Consultando o Balanço Social do Município de Oeiras no ano 2017, é possível saber que 59% dos trabalhadores estão classificados como Assistentes Operacionais o que poderá explicar a maior incidência de acidentes nesta categoria. O outro fator que poderá ter impacto são as funções desempenhadas por estes profissionais. Estas são tendencialmente mais físicas, havendo uma maior probabilidade de acidente (Municipio de Oeiras, 2017).

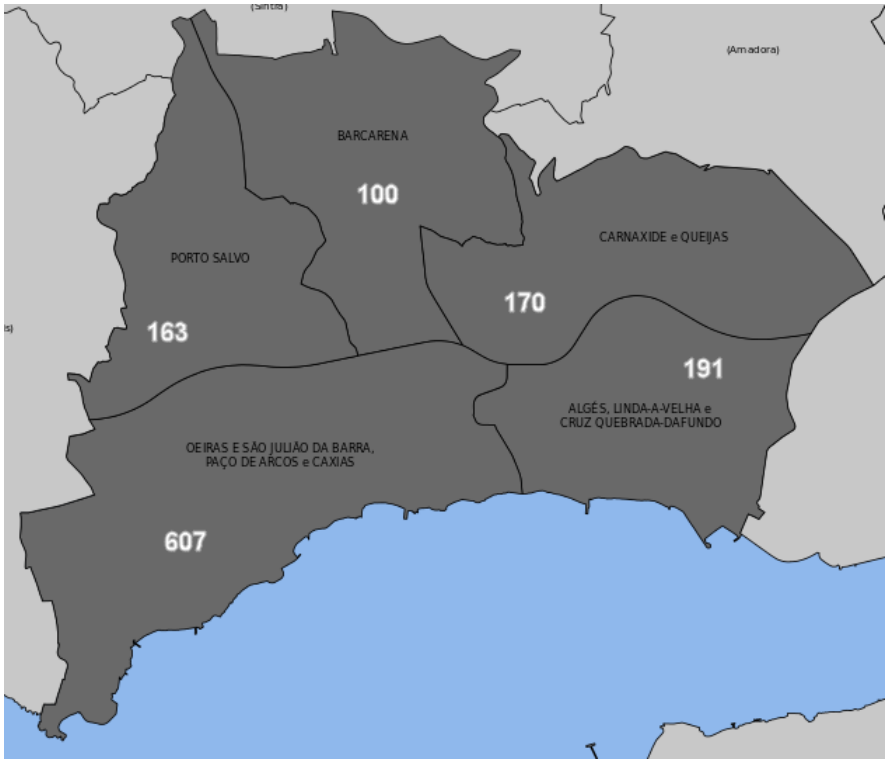

Figura 8 - Distribuição dos acidentes por Freguesia do concelho

O concelho de Oeiras está dividido em cinco freguesias com uma área total de 45,88 km². As sedes municipais encontram-se maioritariamente na freguesia de Oeiras e São Julião da Barra, Paço de Arcos e Caxias, o que poderá explicar a maior ocorrência de acidentes de trabalho neste território como observado na figura 8 (Municipio de Oeiras, 2013).

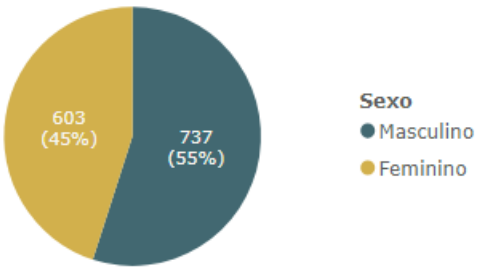

Figura 9 - Distribuição dos acidentes por Sexo

De acordo com o balanço social, 60% dos trabalhadores do município são mulheres, no entanto, através da figura 9, podemos observar que os acidentes ocorrem com maior incidência no sexo masculino. Não foram encontrados dados plausíveis que expliquem esta discrepância.

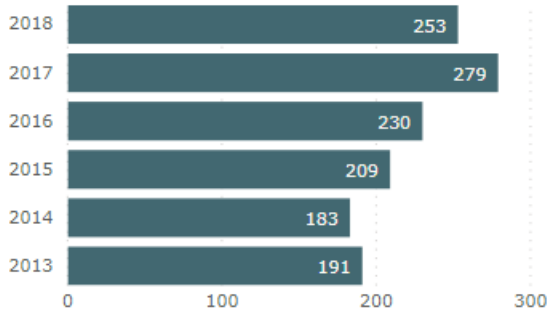

Figura 10 - Distribuição dos acidentes por Ano

Na figura 10, é possível observar que o número de acidentes tem vindo a aumentar de 2013 a 2017, tendo em 2018 havido uma pequena diminuição.

3.4. MISSING VALUES

A abordagem mais comum para a ausência de valores é simplesmente excluir da amostra os casos que contenham variáveis que apresentem valores omissos e analisar os dados restantes. Essa abordagem é conhecida como *listwise deletion* (Kang, 2013). Este método tem como desvantagens a redução do poder estatístico, pois reduz o número de casos estudados e não utiliza toda a informação fornecida (Humphries, 2013). De forma a maximizar a informação, é utilizado por vezes o método *Pairwise Deletion,* pois este preserva os casos que contenham variáveis com valores omissos, eliminando apenas quando o dado específico é necessário para testar uma suposição (Kang, 2013).

Além disso, outra forma de tratar a ausência de valores é usar métodos de imputação. Estes envolvem, por exemplo, um processo de substituição de dados por valores como a média ou mediana (Shrive, Stuart, Quan & Ghali, 2006). As formas mais comuns de imputação são a regressão que, em vez de excluir qualquer caso que possua algum valor ausente, preserva todos os casos, substituindo os dados ausentes por um valor provável estimado por outras informações disponíveis (Kang, 2013).

Para aplicar a melhor técnica a este problema, é necessário compreender os motivos da ausência de informação, pois as técnicas a serem aplicadas variam de acordo com os motivos pelos quais os dados estão ausentes (Soley-Bori, 2013).

Da análise estatística efetuada previamente, podemos observar que a coluna Dias de Ausência contem 196 casos com valor omisso. Foi utilizado o método de imputação substituindo estes valores pela mediana da coluna, *0 dias.* Esta decisão foi tomada pois, no preenchimento de uma ficha manual, os campos de inserção podem ser deixados em branco por serem de valor nulo. Foi realizado o mesmo exercicio para a coluna Total da Despesa, neste caso foram encontrados 194 casos e foram substituidos por €0.

Após a imputação dos passos anteriores, na Tabela 2 é possivel observar a estrutura:

Variável	Nº de Níveis	Nº Valores Omissos
Área	5	0
Assistência Médica	46	0
Categoria	7	0
Concelho	6	70
Intervalo Idade	5	3
Natureza	29	0
No Exercício Laboral?	2	6
Período	3	15
Sexo	2	5
Vínculo	3	1

Tabela 2 - Estrutura da Base de Dados

3.5. CRIAÇÃO DA VARIÁVEL TARGET (FRAUDE)

A modelação preditiva é baseada em técnicas utilizadas para classificação e regressão. Um campo numa base de dados é designado como a variável *target* e o algoritmo produz modelos onde a variável *target* é escrita em função dos outros campos da base de dados, que são pré-identificados como variáveis explicativas (Apte *et al*., 2003).

Foi definido como objetivo a criação de modelo preditivo para a fraude. Nesse sentido, foi necessário criar uma variável binária, Fraude, onde o valor 1 representa a suspeita de fraude e o valor 0 a não fraude. Para o preenchimento desta coluna, foram tidas em consideração as variáveis Natureza e Dias de Ausência. Na Tabela 3, podemos observar o cálculo da mediana e o desvio padrão dos dias de ausência para cada uma das naturezas da lesão do acidente. Na quarta e quinta coluna foram calculados, em dias, o somatório da mediana com o desvio padrão e o somatório da mediana com duas vezes o desvio padrão respetivamente.

Natureza	Mediana (m)	Desvio Padrão (σ)	m+σ	m+2*σ
Alergia	9.00	45.18	54.18	99.36
Algia	0.00	71.39	71.39	142.79
Amputação	0.00	0.00	0.00	0.00
Cervicalgia	0.00	18.37	18.37	36.74
Conjuntivite	4.50	6.95	11.45	18.39
Contusão	10.00	24.69	34.69	59.37
Corpo Estranho	4.00	6.07	10.07	16.13
Corte	2.00	5.43	7.43	12.86
Danos materiais	0.00	0.00	0.00	0.00
Dorsalgia	0.00	16.50	16.50	33.00
Edema	0.00	4.57	4.57	9.15
Entorse	12.00	36.73	48.73	85.46
Esmagamento	194.00	127.81	321.81	449.63
Ferida	0.00	9.19	9.19	18.39
Fratura	65.29	0.00	65.29	65.29

Hematoma	3.00	42.69	45.69	88.37
Hérnia	43.00	4.24	47.24	51.49
Inflamação	0.00	11.66	11.66	23.32
Irritação	5.50	0.71	6.21	6.91
Lesão	1.00	175.81	176.81	352.63
Lesão Muscular	5.00	44.09	49.09	93.18
Lombalgia	5.00	143.65	148.65	292.30
Luxação	15.00	192.96	207.96	400.91
Omalgia	0.00	10.77	10.77	21.53
Outro	0.00	27.91	27.91	55.83
Picada	0.00	2.41	2.41	4.82
Queimadura	0.00	10.50	10.50	21.00
Traumatismo	0.00	91.73	91.73	183.46

Tabela 3 - Mediana e Desvio Padrão dos dias de Ausência por Natureza da Lesão

Com base nas últimas duas colunas da tabela, foram contruídas duas colunas na base de dados:

- Fraude1: acidentes com dia de ausência superiores à m+σ foram considerados suspeitos de fraude (1), restantes foram classificados como não fraude (0);
- Fraude2: acidentes com dia de ausência superiores à m+2*σ foram considerados suspeitos de fraude (1), restantes foram classificados como não fraude (0);

Após o preenchimento destas duas colunas, as mesmas apresentam a seguinte estrutura:

	Fraude1		Fraude2	
	0	1	0	1
Percentagem	90.78%	9.22%	94.87%	5.13%

Tabela 4 - Estrutura variável Fraude1 e Fraude 2

Através da análise da tabela 4, podemos verificar que a percentagem encontrada para a presença de valor positivo na variável Fraude2 não é significativo para a utilização da mesma na restante análise. Pelo que nos passos seguintes será apenas tida em consideração como variável *target* a variável Fraude1.

3.6. SELEÇÃO DE VARIÁVEIS

Se o objetivo é a previsão, é importante saber com que precisão a variável dependente pode ser estimada a partir das variáveis explicativas. Nesse caso, o uso do R^2 pode fazer algum sentido, pois um valor alto geralmente indica um erro de previsão (Moksony, 2014). Com o intuito de realizar uma seleção às variáveis a considerar na construção do modelo, foram analisados os resultados das estatísticas do R^2.

Variável	Nível	Resultado
Área	Nominal	
Assistência Médica	Nominal	
Categoria	Nominal	Rejeitada
Concelho	Nominal	Rejeitada
Dias de Ausência	Intervalar	
Intervalo Idade	Nominal	
Natureza	Nominal	
No Exercício Laboral	Binária	Rejeitada
Período	Nominal	
Sexo	Binária	
Total Despesas	Intervalar	
Zona Atingida	Nominal	

Tabela 5 - Seleção de Variáveis

Após as análises anteriores, iremos percorrer dois caminhos diferentes na construção dos modelos:

1. Tendo em consideração todas as variáveis;
2. Tendo em consideração apenas as variáveis selecionadas através do valor do R^2.

3.7. PARTIÇÃO DE DADOS

A partição dos dados fornece conjuntos de dados mutuamente exclusivos. A base de dados poderá ser partida em dois ou mais conjuntos de dados mutuamente exclusivos que não partilham observações entre si. A partição dos dados de entrada reduz o tempo de computação das execuções de modelagem preliminares (SAS Institue Inc., 2017).

Considerando que a base de dados possui 1.345 observações, a abordagem de partição utilizada foi de 70% para treino, 30% para validação e 0% para teste. Com esta divisão, conseguimos manter mais observações para treinar o modelo e posteriormente obter melhores resultados utilizando a base de dados de validação. Na Tabela 6 é possível observar a distribuição das observações:

	Fraude	
	Treino	Validação
Fraude	85	38
Não Fraude	854	368
TOTAL	939	406

Tabela 6 - Partição da base de dados

3.8. MODELO

Esta fase consiste principalmente na modelagem dos dados modificados anteriores, a fim de procurar automaticamente relações entre eles e prever o resultado desejado. Existem diferentes técnicas disponibilizadas - como Redes Neuronais, Regressão, Árvores de Decisão e Clustering de meios K - e cada uma delas possui vantagens particulares e atua de uma determinada forma, dependendo das situações. O objetivo final é escolher o modelo que, com base em dados históricos, regista alta precisão e é realista ao prever comportamentos futuros da variável de destino.

Como descrito no capítulo anterior, as técnicas de deteção de fraude podem dividir-se em métodos supervisionados e não supervisionados. Neste trabalho, consideraremos apenas métodos supervisionados. Estes algoritmos têm como objetivo principal prever com precisão um valor-alvo específico usando um subconjunto de dados e variáveis para os quais esse valor-alvo já é conhecido. É importante realçar que não há apenas uma solução e improvável que exista alguma completamente correta, pelo que deverá ser um processo interativo. Este implica a aplicação de vários algoritmos que poderão reproduzir resultados diferentes e funcionar de formas distintas. Posteriormente deverá ser realizada uma comparação através de indicadores estatísticos ou de *performance*.

Considerando a revisão da literatura, neste estudo, decidimos implementar três algoritmos diferentes no modelo: Árvores de Decisão, Regressão Logística e Redes Neuronais. O *software* utilizado foi o SAS Enterprise Miner. Este inclui várias vantagens como interface baseada em fluxo com arrastar e soltar e bom manuseio com grandes conjuntos de dados.

Tendo em consideração que as variáveis *Natureza* e *Dias de Ausência* foram utilizadas como base para a construção da variável *target*, *Fraude*, as mesmas foram excluídas dos algoritmos implementados de forma a não obter resultados redundantes.

3.8.1. Árvores de Decisão

As árvores de decisão representam uma segmentação supervisionada modular de uma fonte de dados definida criada pela aplicação de uma série de regras que podem ser determinadas empiricamente e pelos utilizadores. Um nó com todos os seus sucessores forma uma ramificação do nó que o criou. Os nós finais são chamados de folhas (Iain L J Brown, 2014).

Para a utilização do nó Árvore de Decisão no SAS, este requer pelo menos uma variável *target* (resposta) e pelo menos uma variável de entrada (explicativa independente). Múltiplas variáveis de destino, múltiplas variáveis de entrada, múltiplas variáveis de custo e uma única variável de frequência são permitidas para o nó Árvore de Decisão (SAS Institue Inc., 2017).

A folha final contendo uma observação é o seu valor preditivo. Ao contrário das redes neuronais e da regressão logística, as árvores de decisão não funcionam com dados de intervalo. As árvores de decisão trabalham com variáveis nominais que têm mais de dois resultados possíveis e com variáveis ordinais. Valores ausentes podem ser utilizados na criação de regras *if-then*. Portanto,

a imputação não é necessária para as árvores de decisão, embora possa ser utilizada (Cerrito & SAS Institute, 2006).

Na implementação deste modelo foram utilizadas as propriedades *default* do SAS Entreprise Miner para o nódulo da árvore de decisão.

3.8.2. Regressão Logística

O nódulo de Regressão Logística é utilizado para ajustar os modelos de regressão a um conjunto de dados predecessor em um fluxo de processo do SAS Enterprise Miner. A regressão linear tenta prever o valor de um alvo como uma função linear de uma ou mais entradas independentes. A regressão logística tenta prever a probabilidade de um alvo binário ou ordinal adquirir o evento de interesse em função de uma ou mais entradas independentes (SAS Institue Inc., 2017)

Em complemento da regressão simples, poderão ser utilizadas procedimentos de seleção de variáveis como o método de seleção *Forward* e *Backward*. Estes procedimentos são geralmente utilizados para realizar uma triagem inicial das variáveis candidatas (Modeling & Selection, 2008).

Modelo de Seleção – Nenhum: todas as entradas são usadas para ajustar o modo.

Modelo de Seleção – *Forward*
O modelo começa com nenhuma variável selecionada. Posteriormente é selecionada etapa por etapa a variável que possui o R-quadrado mais alto. O processo para de adicionar variáveis quando nenhuma das variáveis restantes for significativa (Modeling & Selection, 2008).

Modelo de Seleção *Backward*
O modelo começa com todas as variáveis selecionadas. Posteriormente, é removida etapa por etapa a variável que possui o R-quadrado mais baixo. O processo continua até que que nenhuma variável não significativa não permaneça (Modeling & Selection, 2008).

Critério de seleção – se for escolhido um dos métodos de seleção mencionados anteriormente, também deverá ser definida a propriedade *critérios de seleção* para especificar o critério de escolha do modelo final. No caso desta análise, foi escolhido o critério padrão que usa profit/loss com dados de treino, profit/loss com dados de validação e nenhum com dados brutos ou de teste (SAS Institue Inc., 2017).

Na implementação deste modelo, foram utilizadas as propriedades *default* do SAS Entreprise Miner para o nódulo de regressão logística, aplicando os 3 modelos de seleção referidos anteriormente.

3.8.3. Redes Neuronais

As redes neuronais (RN) são representações matemáticas modeladas da funcionalidade do cérebro humano. O benefício das RN é a sua flexibilidade na modelação de praticamente qualquer associação não linear entre variáveis de entrada e a variável *target*. As RN mais

utilizadas são a Multilayer Perceptron (MLP), que é composta por uma camada de entrada (que consiste em neurónios para todas as variáveis de entrada), uma camada oculta (que consiste em qualquer número de neurónios ocultos) e uma camada de saída. Cada neurónio processa as entradas e transmite o valor de saída para os neurónios da camada subsequente. Cada uma dessas conexões entre neurónios recebe um peso durante o treino (Iain L J Brown, 2014). Como resultado não existe um modelo ou equação definida e o modelo não é apresentado num formato conciso, como no caso da regressão logística (Cerrito & SAS Institute., 2006).

Na implementação deste modelo, foram utilizadas as propriedades *default* do SAS Entreprise Miner para o nódulo de redes neuronais.

3.9. COMPARAÇÃO DE MODELOS

A comparação de modelos permite perceber o desempenho de modelos concorrentes utilizando vários critérios (SAS Institue Inc., 2017). Algumas das técnicas utilizadas para avaliar o desempenho dos algoritmos utilizados no modelo são: *relative operating characteristic curve* (ROC), *misclassification rate* e a resposta capturada.

A curva ROC é definida através de um gráfico que tem como eixo y a sensibilidade do teste e como eixo do x a especificidade ou a taxa de falso-positivo do teste. Ambos os valores atingem o seu máximo no valor 1. A área debaixo da curva é uma medida que combina a sensibilidade e especificidade. Esta avalia o desempenho geral de um teste e é interpretada no valor médio da sensibilidade para todos os valores possíveis de especificidade. Pode assumir qualquer valor entre 0 e 1, pois os eixos x e y têm valores que variam de 0 a 1. Quanto mais próxima estiver de 1, melhor o desempenho do teste. Um teste com um valor igual a 1 é aquele que é completamente preciso (Park, Goo & Jo, 2004).

A *misclassification rate,* taxa de classificação incorreta, é definida como a percentagem de exemplos de treino e teste erradamente classificados de um determinado conjunto de dados (Raju & Schumacker, 2016). Esta taxa é definida pela equação abaixo:

$$\text{Misclassification rate} = \frac{\text{False positives+false negatives}}{\text{Total instances}}$$

Figura 11 - Equação da Taxa de Classificação Incorreta
Fonte: Misclassification rate (Baughman, Liu, & Engi-, 1995)

Na figura 11 é possível observar que a equação representa a proporção de falsas previsões relativamente ao total de acidentes. O objetivo é atingir o valor mínimo possível para esta estatística, pois indicará que o modelo obtido é mais preciso (Baughman *et al*., 1995).

Resposta capturada é uma estatística que escolhe o modelo com os maiores valores de respostas capturadas num intervalo de decil ou semidecil (Forum, 2015). Nesse sentido, esta estatística considerará o modelo com o valor mais alto. Assim, e para calcular esta estatística, considera-se

que a resposta capturada é o número de respostas em cada decil pelo número total de respostas do evento de destino - neste caso, uma fraude.

4. RESULTADOS E DISCUSSÃO

Após concluir a fase de modelagem, enquanto diferentes premissas foram testadas, o Nó de Comparação do Modelo foi usado para avaliar que modelo teve melhor desempenho ao prever. Nesse sentido, todas as combinações de resultados foram comparadas para atingir esse objetivo.

- **Todas as variáveis:**

Considerando primeiro a abordagem de manter todas as variáveis, a Tabela 7 ilustra as combinações testadas para alcançar os valores desejáveis para cada uma das estatísticas selecionadas para o modelo preditivo. Considerando as estatísticas mencionadas no capítulo anterior, os resultados considerando o melhor modelo para cada estatística são os seguintes:

Critério de Seleção	Modelo Escolhido	Valor
Misclassification Rate	Redes Neuronais	0.09113
ROC Index	Regressão - Forward	0.74400
Resposta Capturada	Regressão - Backward	19.29825

Tabela 7 - Todas as variáveis - Resultados

- **Com seleção de variáveis:**

A abordagem a seguir foi removendo algumas variáveis, e as combinações testadas para atingir os valores desejáveis para cada uma das estatísticas selecionadas para o modelo preditivo estão representadas na Tabela 8. Considerando as estatísticas mencionadas no capítulo anterior, os resultados considerando o melhor modelo para cada estatística são os seguintes:

Critério de Seleção	Modelo Escolhido	Valor
Misclassification Rate	Regressão - Forward	0.09076
ROC Index	Regressão - Forward	0.72900
Resposta Capturada	Redes Neuronais	12.72727

Tabela 8 - Com seleção variáveis – Resultados

Adicionalmente à analise anteriormente realizada, foi testado outro modelo. A base de dados foi segmentada pelas categorias da variável *Natureza.* A categoria que contém mais observações é o **Traumatismo**, 433 acidentes desta natureza, tendo sido por isso a categoria escolhida para a realização de um modelo modelo especifico.

Tendo como base todos os acidentes ocorridos nos último cinco anos na CMO com a natureza traumatismo, foram aplicados os mesmos algoritmos que foram aplicados no modelo geral. O modelo que obteve uma taxa de classificação incorreta mais baixa foi a regressão logistica pelo que é o modelo seleccionado e apresentado nos resuldados abaixos.

Variável	p-value
Área	<0.0001
Assistência Médica	<0.0001
Categoria	<0.0001
Concelho	0.9270
Intervalo Idade	0.1723
No Exercício Laboral?	0.4268
Período	0.0010
Sexo	0.4096
Zona Atingida	0.3836

Tabela 9 - Resultados com 95% de confiança

Pela tabela 9 pode-se perceber que as variáveis que detêm maior peso na variável Target são a *Área, Assistência Médica e Categoria*. Dentro destas 3 variáveis, as categorias que apresentam um maior valor de β, isto é, têm um maior impacto na suspeita de existência de fraude na equação contruída, são:

- Área:
    1. Limpeza/Higiene - β = 3.14;
    2. Operacional - β = 2.40;
- Assistência Médica:
    1. Hospital CUF Cascais - β = 12.99;
- Categoria:
    1. Assistente Técnico - β = 7.25.

Outras categorias têm um impacto contrário, como é o caso da área de segurança que tem um β negativo de -10, diminuindo a probabilidade de suspeita de fraude, caso o acidente tenha ocorrido a um funcionário desta área.

Os resultados completos deste modelo de regressão logistica encontram-se no anexo II.

5. CONCLUSÕES

O principal objetivo desta dissertação de mestrado foi avaliar a possibilidade de construir um modelo analítico que, com base no histórico, previsse a ocorrência de fraude nos acidentes de trabalho na CMO. Para atingir esse objetivo, numa fase inicial, foi realizada uma revisão completa do tema, acidentes de trabalho e fraude e analisados os potenciais modelos a testar. Posteriormente, foi recolhido o histórico dos acidentes ocorridos nos últimos 5 anos. Esta informação foi fornecida pela Unidade de Segurança e Saúde do Trabalho.

Foi iniciado o processo de limpeza e transformação dos dados. Com uma base de dados pronta a ser utilizada, os modelos estudados foram construídos tendo em consideração as 10 variáveis: Área, Assistência Médica, Categoria, Concelho, Intervalo Idade, No Exercício Laboral?, Período, Sexo, Total Despesas e Zona Atingida. No diagrama disponível no anexo III, é possível verificar que foram realizadas duas análises separadas, uma utilizando o nódulo de seleção de variáveis (1) e a outra análise não utilizando esse nódulo (2). Em ambos os casos, os resultados ficaram aquém do esperado. Isto é, apesar da taxa de classificação incorreta tomar valores muitos próximos de 0, o valor da resposta capturada em todos os algoritmos testados fica abaixo dos 20%.

Ambos os métodos são relevantes para este estudo, pois referem-se a diferentes abordagens. Nesse sentido, é possível escolher o melhor considerando as informações disponíveis - com mais ou menos variáveis. Comparando os algoritmos apenas pela taxa de classificação incorreta é possível observar que o modelo com melhores resultados na análise (1) é a regressão logística com o método de seleção *Forward*. Neste modelo, as variáveis mais explicativas para a fraude são: Total Despesas, Zona Atingida e a Área. A partir da tabela de classificação, é possível concluir que este modelo classificou erradamente 551 de 606 observações, demonstrando que não tem a eficácia pretendida. Já na análise (2), o modelo com melhores resultados foi o das redes neuronais. Neste modelo a percentagem de observações erradamente classificados é de 91%, pelo que também não o podemos considerar um modelo eficaz.

Esta conclusão levou-nos à construção um modelo segregando os dados por natureza de acidente. Este modelo foi construído através da regressão logística pois este método apresentou um valor de resposta capturada perto dos 30%. Como resultado aferiu-se que as variáveis *Área, Assistência Médica e Categoria* são as que mais influenciam positivamente a suspeita de fraude. Não foram construídos os modelos para todos os tipos de Natureza pelo facto de algumas categorias conterem poucas observações.

Resumindo, nesta dissertação foram contruídos 3 modelos, um modelo geral com seleção de variáveis (Regressão – *Forward*), um modelo geral sem seleção de variáveis (Redes Neuronais) e um modelo para natureza traumatismo (Regressão Logística).

Observando a eficácia e precisão dos mesmos podemos concluir que a construção de um modelo preditivo para a fraude poderá ter melhores resultados agrupando os acidentes por Natureza.

Por fim, este projeto acompanha a necessidade de modelos atualizados para evitar fraudes. De estudos anteriores, são conhecidos os principais riscos de fraude numa empresa. No entanto, existe claramente uma lacuna na deteção de fraudes em acidentes de trabalho. Não existem muitos estudos considerando esta área específica, logo o estudo realizado teve como objetivo preencher essa lacuna fornecendo informações relevantes que podem ser utilizadas noutros ensaios deste campo. Assim, este projeto também se esforça para fornecer novas perspetivas e metodologias para entender as razões pelas quais a fraude ocorre e tentar melhorar os métodos para evitar esse problema.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

A principal limitação encontrada ao desenvolver o modelo apresentado nesta dissertação foi a baixa quantidade e qualidade dos dados. Como a informação do acidente é recolhida através do preenchimento de uma ficha com campos de texto aberto, a informação não estava normalizada. O *data mining* é uma técnica que depende fortemente dessas duas características. Considerando que a base de dados inicialmente continha 1393 observações e 39 variáveis e, após a aplicação das primeiras etapas da metodologia para limpeza e transformação, apenas ficaram 1345 observações e 8 variáveis que foram utilizadas nas etapas seguintes.

Outra limitação encontrada foi a inexistência da variável fraude, o que levou à sua construção baseada em revisão de literatura e aconselhamento profissional, mas que poderá ter tido impacto os resultados obtidos. Isto significa que a volatilidade desta variável dificulta a construção de um modelo mais robusto para os casos de fraude.

A última limitação encontrada foi a falta de documentação sobre fraudes nos acidentes de trabalho, pelo que a revisão da literatura apenas teve como base uma fonte.

Como referência e sugestão para trabalhos futuros, propõe-se consolidar uma base de dados maior e com os casos de fraude previamente identificados. Além disso, também seria interessante desenvolver diferentes modelos focados nos diferentes tipos de acidentes, por exemplo por exercício profissional. Ao dividir a análise, novos e atraentes comportamentos podem ser descobertos.

7. BIBLIOGRAFIA

ABI. (2011). *ABI comments on the Basel Committee on Banking Supervision consultative document : " Sound Practices for the Management and Supervision of*. (February).

Apte, C. V., Hong, S. J., Natarajan, R., Pednault, E. P. D., Tipu, F. A., & Weiss, S. M. (2003). Data-intensive analytics for predictive modeling. *IBM Journal of Research and Development*, *47*(1), 17–23. https://doi.org/10.1147/rd.471.0017

Baesens, B., Vlasselaer, V. Van, & Verbeke, W. (2015). Fraud: Detection, Prevention, and Analytics! *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques*, 1–36. https://doi.org/10.1002/9781119146841.ch1

Basel Committee on Banking Supervision. (2001). Consultative Document: Operational Risk. *Basel Capital Accord*, (May).

Baughman, D. R., Liu, Y. A., & Engi-, C. (1995). *Misclassification Rate Classification : Fault Diagnosis and Fea- ture Categorization Is environmental tobacco smoke a risk factor for lung cancer ?*

Beatriz, M. (2019). O conceito de acidente de trabalho: conexão com a relação laboral. *Imprensa Da Universidade de Coimbra*. Retrieved from http://hdl.handle.net/10316.2/42159

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, *17*(3), 235–255. https://doi.org/10.1214/ss/1042727940

Boniface, D. R., & Boniface, D. R. (2019). Multiple regression. *Experiment Design and Statistical Methods*, 130–142. https://doi.org/10.1201/9780203756423-11

Cerrito, P. B., & SAS Institute. (2006). *Introduction to data mining using SAS Enterprise Miner*.

Cilimkovic, M. (2010). Neural Networks and Back Propagation Algorithm. *Fett.Tu-Sofia.Bg*, 3–7. Retrieved from http://fett.tu-sofia.bg/et/2006/ET2006 BOOK 1/Circuits and Systems/173 Paper-V_Skorpil.pdf

Cross, S. S., Harrison, R. F., & Kennedy, R. L. (1995). Introduction to neural networks. In *The Lancet* (Vol. 346). https://doi.org/10.1016/S0140-6736(95)91746-2

Denominations, C., & Church, R. C. (n.d.). *The Catholic Encyclopedia , Volume 1 : Aachen-As-Publisher*

Diário da República n.º 77/2017, Série II de 2017-04-19. Aviso n.º 4191/2017, 4.1.

Estado Português. (2009). Lei nº 98/2009, de 4 de Setembro, que Regulamenta o regime de reparação de acidentes de trabalho e de doenças profissionais, incluindo a reabilitação e reintegração profissionais, nos termos do artigo 284.º do Código do Trabalho, aprovado pela Lei n.º 7/2. *Diário Da República*, *1ª Série*(Nº 172), 5894–5920.

European Banking Authority. (2016). Operational risk. Retrieved from http://www.eba.europa.eu/regulation-and-policy/operational-risk, consultado em 25/06/2019

European Parliament and European Council. (2009). Directive 2009/138/EC (Solvency II). *Official Journal of European Union*, *L335*, L335/1-129. https://doi.org/10.3000/17252555.L_2009.335.eng

Fahmi, M., Hamdy, A., & Nagati, K. (2016). Data Mining Techniques for Credit Card Fraud Detection : Empirical Study. *Sustainable Vital Technologies in Engineering & Informatics*, (2015), 1–9.

Forum, G. (2015). *Introduction to Data Mining SAS ® Global Forum 2015 Handout*.

GEP. (2015). Acidentes de trabalho 2013. *Gabinete de Estratégia e Planeamento, MSESS*. https://doi.org/10.1590/S0102-311X2003000200015

Glover, H., & Flagg, J. (1999). *Fraud Detection and Prevetion.doc*. Retrieved from https://chapters.theiia.org/ottawa/Documents/Fraud_Detection_and_Prevetion.pdf

Gupta, S. (2015). A Regression Modeling Technique on Data Mining. *International Journal of Computer Applications*, *116*(9), 27–29. https://doi.org/10.5120/20365-2570

Hawlova, K. (2013). Fraud detection tools. *Journal of Systems Integration*, *4*(4), 10–18. https://doi.org/10.20470/jsi.v4i4.173

Henriques, C. (2011). *Análise de Regressão Linear Simples e Múltipla*. 1–44.

Hermann, Dr, M., & Guamieri, M. (1996). Public Health Then and Now Accidents and Acts of God: A History of the Terms. *American Journal of Public Health*, *86*(1), 101–107. https://doi.org/10.2105/AJPH.86.1.101

Huang, H. (2013). Using Artificial Neural Networks to Establish a Customer-cancellation Prediction Model. *Przegląd Elektrotechniczny*, (January 2013).

Humphries, M. (2013). Missing Data & How to Deal: An overview of missing data. *Population Research Center*, 45. Retrieved from http://www.texaslonghornsl.com/cola/centers/prc/_files/cs/Missing-Data.pdf

Iain L J Brown, P. (2014). *Developing Credit Risk Models Using SAS Enterprise Miner and SAS/STAT: Theory and Application*.

Kaļiņina, D., & Voronova, I. (2014). Risk Management Improvement under the Solvency II Framework. *Economics and Business*, *24*(24), 29. https://doi.org/10.7250/eb.2013.004

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data Mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, *32*(4), 995–1003. https://doi.org/10.1016/j.eswa.2006.02.016

Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004). Survey of fraud detection techniques. *Conference Proceeding - IEEE International Conference on Networking, Sensing and Control*, *2*, 749–754. https://doi.org/10.1109/icnsc.2004.1297040

Mackenzie, U. P., Kimura, H., Kerr, R. B., Mackenzie, U. P., & Lima, F. G. (2010). *OPERATIONAL RISK MANAGEMENT IN NON-FINANCIAL INSTITUTIONS : CASE*. (January).

Marel, Annual Report 2017 – Risk Management
http://ar2017.marel.com/responsible-growth/risk-management, consultado em 20/02/2020

Mendes, J. M. (2015). Ulrich Beck : a imanência do social e a sociedade do risco por José Manuel Mendes. *Instituto de Ciências Sociais Da Universidade de Lisboa.*

Modeling, S., & Selection, M. (2008). *Stepwise Regression Stepwise Regression*. 1–33.

Moksony, F. (2014). Small is beautiful, Interpretation of R 2 in social research. *Szociologiai Szemle*, *Special is*(January 1999), 130–138.

Municipio de Oeiras. (2013). *Oeiras Factos e Números*. 256.

Municipio de Oeiras. (2017). *Balanço Social 2017 - Município de Oeiras*. (c).

Município de Oeiras, 2017. Página da Unidade de Segurança e Saúde no Trabalho
http://www.cm-oeiras.pt/pt/municipio/camara-municipal/organograma/Paginas/usst.aspx, consultado em 12/01/2019

Navarro, A. F. (2017). O Triângulo dos acidentes do trabalho : Uma evolução histórica. *Researchgate*, (May). Retrieved from https://www.researchgate.net/publication/316997558_O_Triangulo_dos_acidentes_do_trabalho_Uma_evolucao_historica

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Park, S. H., Goo, J. M., & Jo, C. H. (2004). Receiver operating characteristic (ROC) curve: Practical review for radiologists. *Korean Journal of Radiology*, *5*(1), 11–18. https://doi.org/10.3348/kjr.2004.5.1.11

Patel, S. (2010). *Quantifying operational risk Appendix Insurance operational risk taxonomy : Solvency II / Basel II Level 1 , Basel II Level 2 , ORIC Level 3*. (May).

Petraşcu, D., & Tieanu, A. (2014). The Role of Internal Audit in Fraud Prevention and Detection. *Procedia Economics and Finance*, *16*(December 2014), 489–497. https://doi.org/10.1016/s2212-5671(14)00829-6

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). *A Comprehensive Survey of Data Mining-based Fraud Detection Research*. https://doi.org/10.1016/j.chb.2012.01.002

Pimenta, C. (2009). Esboço de Quantificação da Fraude em Portugal. In *Working Papers - OBEGEF*. Retrieved from http://www.gestaodefraude.eu/index.php?option=com_content&view=article&id=72&Itemid=76

PMI. (2008). *Guide to Project Management Body of Knowledge. 4th ed. USA: PMI; 2008. ISBN 978-1-933890-51-7*.

PorData, 2017. População residente: total e por grandes grupos etários https://www.pordata.pt/Municipios/Popula%C3%A7%C3%A3o+residente+total+e+por+grandes+grupos+et%C3%A1rios-390, consultado em 13/01/2019

Raju, D., & Schumacker, R. (2016). Comparing Data Mining Models in Academic Analytics. *International Journal of Knowledge-Based Organizations*, *6*(2), 38–54. https://doi.org/10.4018/ijkbo.2016040103

SAS Institue Inc. (2017). *SAS ® Enterprise Miner ™ 14.2: Reference Help*. 321–327.

Shao, J., & Pound, C. J. (1999). Extracting business rules from information systems. *BT Technology Journal*, *17*(4), 179–186. https://doi.org/10.1023/A:1009619730683

Shrive, F. M., Stuart, H., Quan, H., & Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Medical Research Methodology*, *6*, 1–10. https://doi.org/10.1186/1471-2288-6-57

Silva, A., Korzenowski, A. L., & Vaccaro, G. L. R. (2014). Uma aplicação da lei de Benford na identificação de padrões estatisticamente assinaláveis de suspeitas de fraude por lavagem de dinheiro. *Espacios*, *35*(7).

Soley-Bori, M. (2013). *Dealing with missing data: Key assumptions and methods for applied analysis*. (4). Retrieved from https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

State Insurance Regulatory Authority (SIRA) https://www.sira.nsw.gov.au/fraud-and-regulation/preventing-fraud/workers-compensation-fraud, consultado em 20/02/2020

Valdez, P. (2019). Circadian rhythms in attention. *Yale Journal of Biology and Medicine*, *92*(1), 81–92.

Witten, I. H., Frank, E., & Hall, M. a. (2011). Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). In *Complementary literature None*. Retrieved from http://books.google.com/books?id=bDtLM8CODsQC&pgis=1

8. ANEXOS

8.1. ANEXO I - FICHA DE COLABORADOR

Modelo 1

Oeiras
Marca o ritmo

PARTICIPAÇÃO DE ACIDENTE DE TRABALHO

QUALIFICAÇÃO E DESPACHO AUTORIZADOR DE DESPESAS

Face aos elementos constantes da participação e aos fornecidos pelo competente serviço de saúde e tendo sido qualificado como acidente de trabalho por despacho do(a) Senhor(a) Presidente, autorizo as despesas dele resultantes.

_____, ___/___/_____

_____

IDENTIFICACÃO DO SERVICO OU ORGANISMO

Designação: **Município de Oeiras**

Morada: **Largo Marquês de Pombal, 2784-501 Oeiras**

Unidade orgânica onde o trabalhador exerce funções _____

_____ Telef: ☐☐☐☐☐☐☐☐☐ Fax: ☐☐☐☐☐☐☐☐☐

IDENTIFICAÇÃO DO TRABALHADOR

Nome _____

_____ Nº.mec. ☐☐☐☐

Data de Nasc. ☐☐ ☐☐ ☐☐☐☐ Nº. Contribuinte ☐☐☐☐☐☐☐☐☐

Nacionalidade _____ Bilhete de Identidade ☐☐☐☐☐☐☐☐

Morada _____

Cod.Postal ☐☐☐☐ ☐☐☐ _____Telef ☐☐☐☐☐☐☐☐☐

Categoria de _____ Área funcional de _____

Adaptado do Decreto-Lei n.º 503/99 de 20/11 – Revisto em 09/01/19

Acidente ☐ Incidente ☐ Acontecimento perigoso ☐

Data da ocorrência ☐☐ ☐☐ ☐☐☐☐ Hora ☐☐ h ☐☐ m, Local _____

_____

Se o acidente não ocorreu no estabelecimento, indique:

Em serviço no exterior ☐

No trajeto residência/trabalho ou vice-versa ☐

Quem prestou os primeiros socorros _____ Ficou hospitalizado? Sim ☐ Não ☐

Foi acidente de viação? Sim ☐ Não ☐ Se Sim, e se o acidente foi da responsabilidade de terceiros indique: Nome do responsável _____

Morada _____

Matrícula do veículo _____ Nº de Apólice _____ Seguradora _____

Se houve intervenção das Autoridades, especifique: _____

DESCRIÇÃO DETALHADA DA OCORRÊNCIA

_____

_____

_____

_____

_____

Tipo de lesão _____ Parte do corpo _____

Testemunhas (indicação não obrigatória) _____

_____

Data ____/____/_____ O DECLARANTE _____

Nota: Participar ao superior hierárquico no prazo de 2 dias úteis após a ocorrência

AQUANDO DA OCORRÊNCIA

Que tarefa efetuava? _____

Tarefa habitualmente exercida? Sim ☐ Não ☐

Utilizava máquinas ou equipamentos? _____ Se sim, quais? _____

_____

Que Equipamentos de Proteção Individual (EPI's) usava? _____

_____

Que medidas deverão ser tomadas para evitar acidentes similares no futuro? _____

_____

_____

O SUPERIOR HIERÁRQUICO _____ Data ____/____/_____

Nota: Remeter à USST no prazo de 1 dia útil após a data em que teve conhecimento

8.2. ANEXO II - MODELO DE REGRESSÃO – TRAUMATISMO

Variáveis Significativas:

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Area	3	116.2963	<.0001
Assist_ncia_M_dica	5	1882.4898	<.0001
Categoria	3	35.7105	<.0001
Concelho	5	1.3752	0.9270
Intervalo_Idade	4	6.3833	0.1723
No_Exercicio_Laboral_	1	0.6315	0.4268
Periodo	2	13.8390	0.0010
Sexo	1	0.6799	0.4096
Vinculo	0	0.0000	.
Zona_atingida	4	4.1689	0.3836

Betas para a Area, Assistência Médica e Categoria:

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-24.8231	151.2	0.03	0.8696	0.000
Area	Administrativo	1	1.2029	0.8938	1.81	0.1784	3.330
Area	Limpeza / Higiene	1	3.1437	0.5225	36.20	<.0001	23.189
Area	Operacional	1	2.4043	0.5821	17.06	<.0001	11.071
Area	Segurança	1	-10.0049	.	.	.	0.000
Assist_ncia_M_dica	ACES Oeiras	1	-4.9976	.	.	.	0.007
Assist_ncia_M_dica	Bombeiros	1	-6.5924	.	.	.	0.001
Assist_ncia_M_dica	Centro Hospital Lisboa Ocidental	0	0	.	.	.	.
Assist_ncia_M_dica	Centro Hospitalar Lisboa Ocident	1	7.8694	88.3313	0.01	0.9290	999.000
Assist_ncia_M_dica	Centro Saúde Oeiras	1	-7.0723	.	.	.	0.001
Assist_ncia_M_dica	Centro de Saúde da Costa da Capa	1	9.4616	.	.	.	999.000
Assist_ncia_M_dica	Centro de Saúde de Oeiras	1	-5.0493	.	.	.	0.006
Assist_ncia_M_dica	Clínica Joaquim Chaves Saúde	1	-3.0017	.	.	.	0.050
Assist_ncia_M_dica	Clínica de Todos os Santos	1	-6.9096	.	.	.	0.001
Assist_ncia_M_dica	Hospital CUF Cascais	1	12.9931	88.3485	0.02	0.8831	999.000
Assist_ncia_M_dica	Hospital Fernando da Fonseca	1	11.2071	88.3409	0.02	0.8990	999.000
Assist_ncia_M_dica	Hospital Garcia da Orta	1	-5.6707	.	.	.	0.003
Assist_ncia_M_dica	Hospital da Luz - Clínica de Oei	1	-5.3120	441.6	0.00	0.9904	0.005
Assist_ncia_M_dica	Hospital de Cascais	1	8.6663	88.3371	0.01	0.9218	999.000
Assist_ncia_M_dica	Hospital de Santa Maria	1	8.7093	.	.	.	999.000
Assist_ncia_M_dica	Hospital dos Lusíadas	1	-7.4443	.	.	.	0.001
Assist_ncia_M_dica	Local	1	-5.5176	.	.	.	0.004
Assist_ncia_M_dica	USF Conde de Oeiras	1	-4.3929	.	.	.	0.012
Assist_ncia_M_dica	USF S. Julião Oeiras	1	-5.9913	.	.	.	0.003
Categoria	Agente PM	1	3.9547	820.6	0.00	0.9962	52.182
Categoria	Assistente Operacional	1	5.1113	.	.	.	165.888
Categoria	Assistente Técnico	1	7.2516	1.2135	35.71	<.0001	999.000
Categoria	Informático	1	-8.7871	968.1	0.00	0.9928	0.000

8.3. ANEXO III - DIAGRAMA SAS MINER